# **Employee Attrition Prediction**



**1. What is employee attrition?**

Employee Attrition is a reduction of an employee for a voluntary or involuntarily reasons. Employee Attrition can cost anywhere from $1,500 to $16,650 per agent.

**1.1 WHAT ARE THE COSTS OF EMPLOYEE ATTRITION AND HIGH ATTRITION RATES?**

There are two sides to employee turnover:

1. Positive
2. Negative

***Positive attrition*** occurs when low-performing workers leave voluntarily or are fired.

***Negative attrition*** when top-performing employees who are responsible for driving sales and increasing revenue become demotivated and start looking for the exit. Negative attrition implies a larger, more serious problem within an organization.

**1.2 Reasons:**

* Poor management increases employee turnover & increases the attrition rate.
* Lack of growth and advancement opportunities is a reason for attrition.
* Inaccurate job profiles contribute to job turnover and attrition rates.
* Poor training can cause a high attrition rate.

**1.3 What is HR Analytics?**

HR analytics is the process of collecting and analyzing Human Resource (HR) data in order to improve an organization’s workforce performance.

The process can also be referred to as talent analytics, people analytics, or even workforce analytics.

**1.4 Role of HR Analytics:**

* Collect and analyze past data on turnover to identify trends and patterns indicating why employees quit.
* Collect data on employee behavior, such as productivity and engagement, to better understand the status of current employees.
* Help create a predictive model to better track and flag employees who may fall into the identified pattern associated with employees that have quit.

**2. Problem Definition:**

In this problem, we have used IBM HR Analytics Employee Attrition & Performance Dataset, which was downloaded from Kaggle.

[https://www.kaggle.com/pavansubhasht/ibm-](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset#WA_Fn-UseC_-HR-Employee-Attrition.csv)hr-analytics-attrition-dataset#WA\_Fn-UseC\_-HR-Employee-Attrition.csv

The dataset includes features like Age, Employee Role, Daily Rate, Job Satisfaction, Years at Company, Years in Current Role, etc. data for 1,470 employees with various information about the employees.

For this problem, we will try to study the factors that lead to employee attrition.

Given that we have data on former employees, this is a **standard supervised classification problem** where the label 0 is an active employee, 1 is former employee. In this study, our target variable Attrition is the probability of an employee leaving the company.

We will be using a step-by-step systematic approach using a method that could be used for a variety of ML problems.

This project would fall under what is commonly known as HR Analytics or People Analytics.

*For complete code, please refer to this* [***GitHub repo***](https://github.com/vigneshrajancse/DataScience-Practice/blob/main/Evaluation%20Project/HR%20Analytics.ipynb)***.***

**3. Data Analysis:**

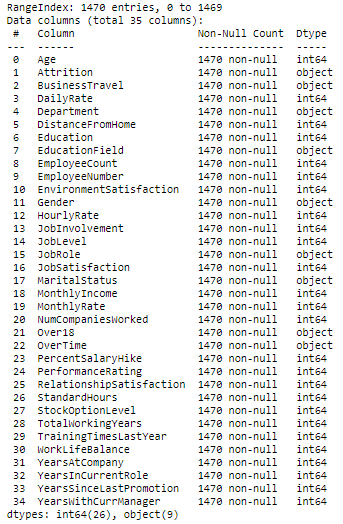
In this problem, we will use [this](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset#WA_Fn-UseC_-HR-Employee-Attrition.csv) dataset to predict when employees are going to quit by understanding the main drivers of employee churn.

**3.1 Reading CSV file:**

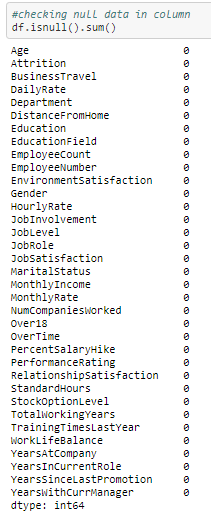
Reading CSV file using pandas, where it has 1470 rows and 35 columns. 

**3.2 Information about dataset:**

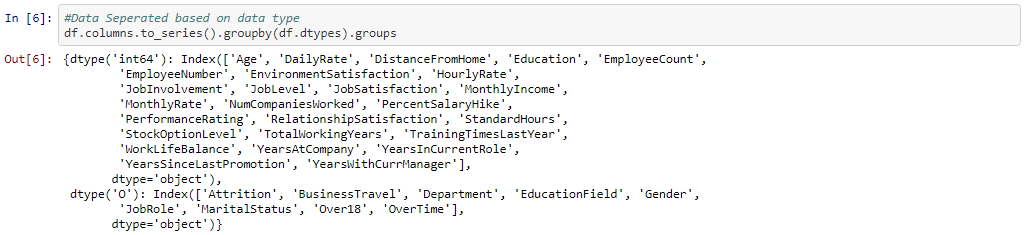
In this dataset there are 35 columns which has numerical (int64) and categorical (object) data type.



We don’t have any missing values in the dataset, and it’s a good thing to start work.



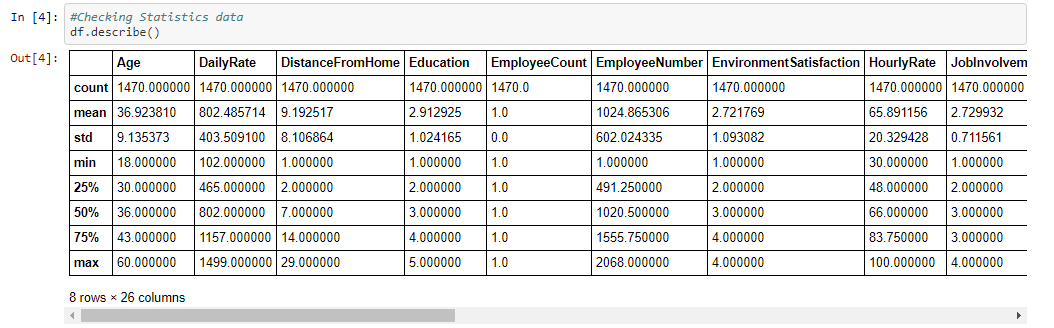
Then we are grouping integer and categorical data type using group by function for columns which separates our dataset columns.



**3.3 Summary Statistics:**

Below describe method shows the statistics of numerical data that is integer or float data type.

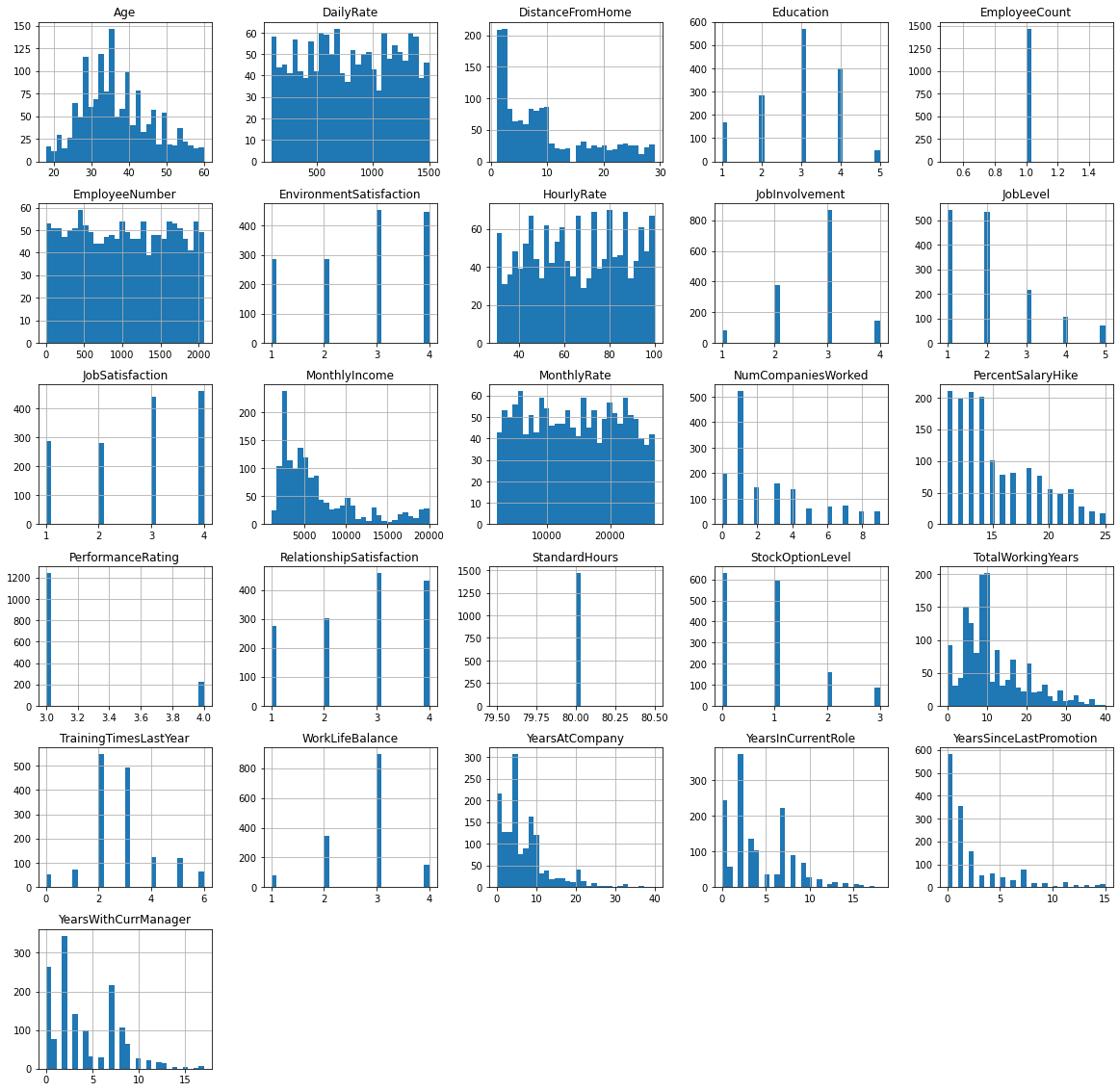
Since we don’t have any null values, so it's showing the same count (1470) for all columns. Then it shows mean, median, standard deviation (std), quartile data.



**4. Visualization of Dataset:**

**4.1 Histogram:**

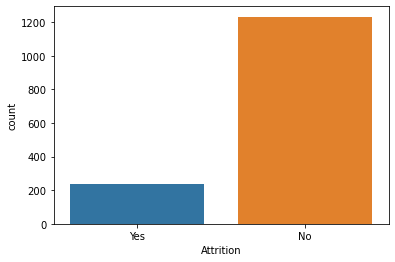
* DistanceFromHome, MonthlyIncome, YearsAtCompany are right skewed.
* There are discrete values like Education, Employee count, Joblevel etc.
* Employee Number is likely to be a unique identifier for employees.
* Age distribution is slightly right-skewed normal distribution between 25 and 45 years old.
* EmployeeCount and StandardHours are constant values for all employees. They’re likely to be redundant features.
* Data transformation methods may be required to approach a normal distribution prior to fitting a model to the data.



## **4.2 Feature distribution by target variable**

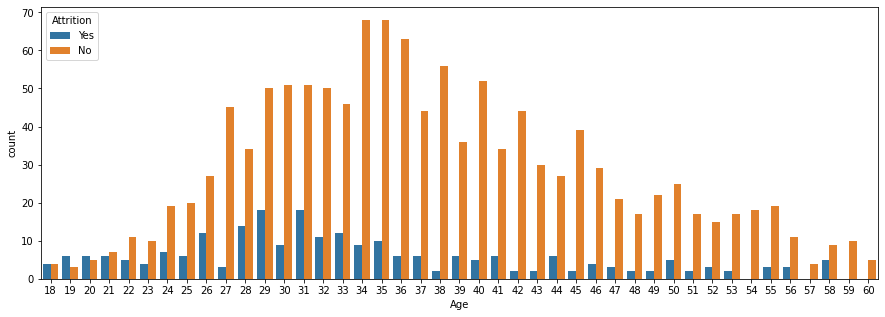
**4.2.1 Attrition:**

In this dataset we have 83 % of No (Active Employee) and 17 % of Yes (Former Employee). Dataset is balanced, so we can predict why there is employee attrition.



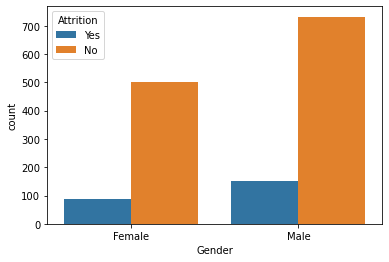
**4.2.2 Age:**

Age distribution shows that between 26 and 35 have the maximum number of former employees and age between 34 and 36 have the maximum number of active employees.



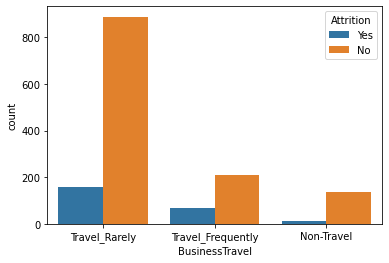
**4.2.3 Gender:**

Gender distribution shows that male has maximum number of former employee than female, and male also have maximum number of active employee than female.



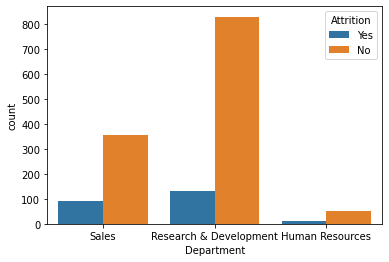
**4.2.4 Business Travel:**

Business Travel distribution shows that largest normalized proportion of Leavers that travel rarely and smallest proportion of Leavers for non- travel category.



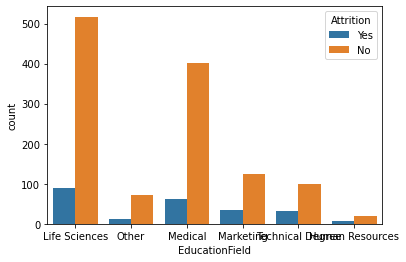
**4.2.5 Department:**

Department distribution shows that Research & Development department shows the largest proportion of former employees than any other department and Human resources department has a small proportion of former employees.



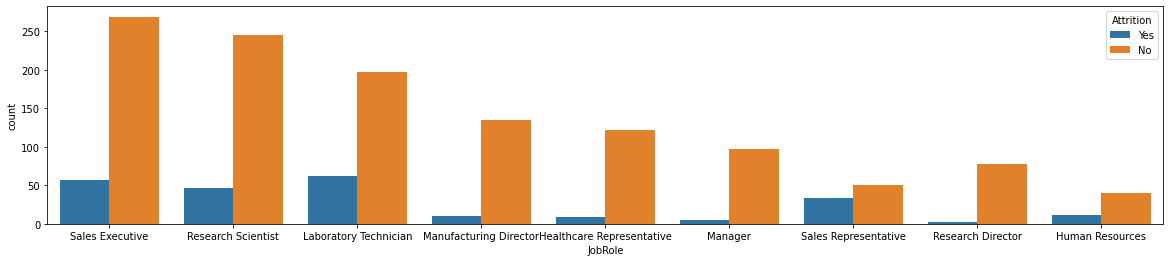
**4.2.6 Education Field:**

Education Field distribution shows that Life Sciences and Medical education field has largest proportion of ex-employees than other education field.



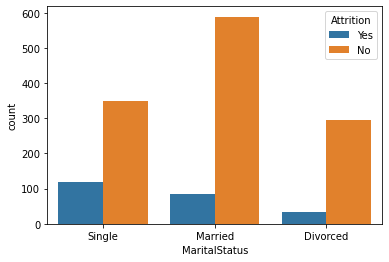
**4.2.7 Job Role:**

Job Role distribution shows that Sales Executive, Research Scientist, laboratory Technician and Sales Representative job roles have largest proportion of ex-employees.



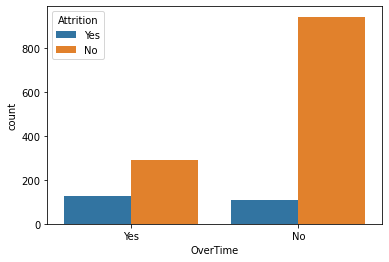
**4.2.8 Marital Status:**

Marital Status distribution shows that those are single have the large proportion of leaving the company than Married & Divorced.



**4.2.9 Overtime:**

Overtime distribution shows that employee who work over time have the largest proportion of leavers than those who did not work over time.

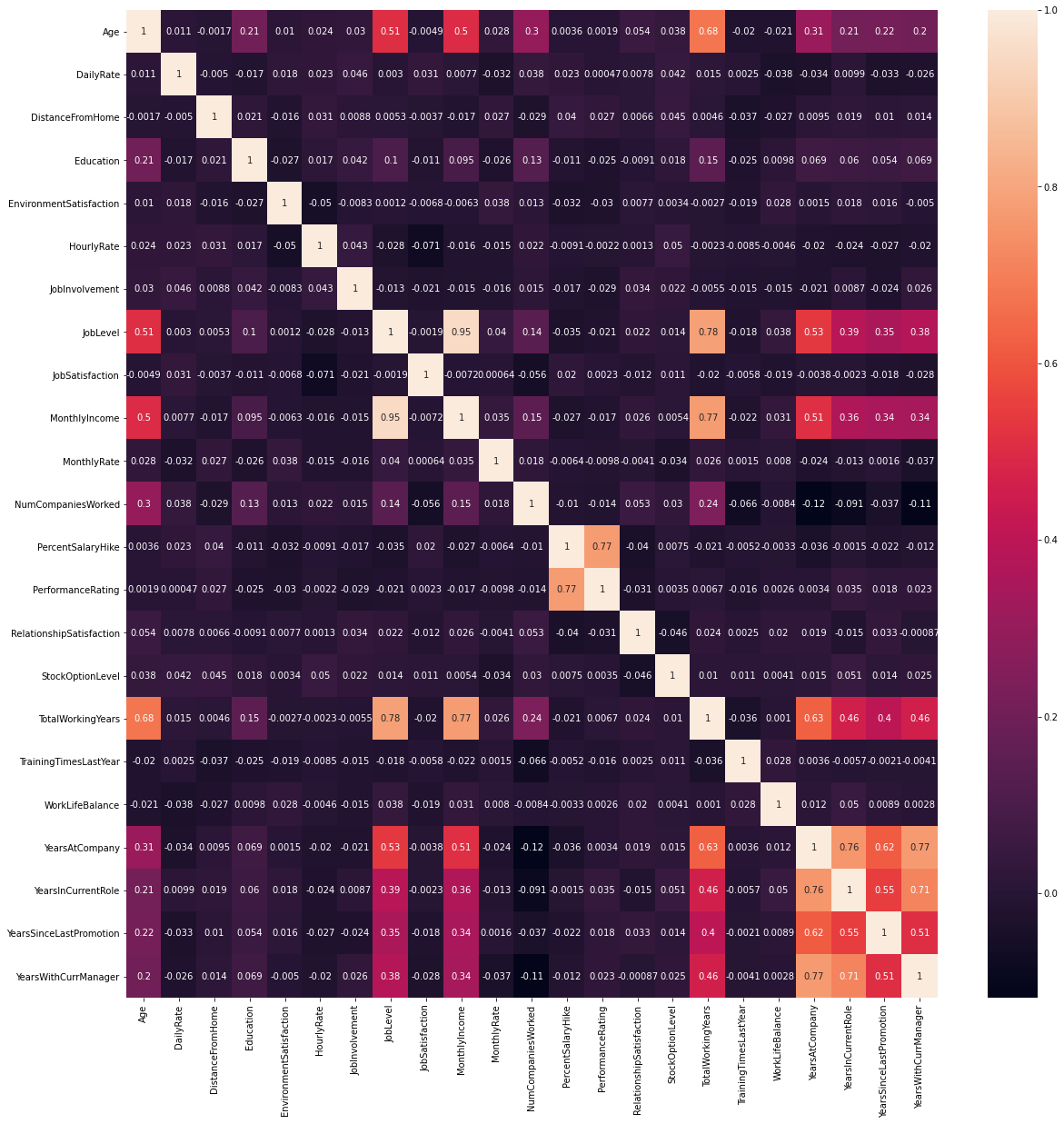


**5. Correlation:**

* Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable.
* It’s a bi-variate analysis measure which describes the association between different variables.

**Positive Correlation:** Two features (variables) can be positively correlated with each other. It means that when the value of one variable increase then the value of the other variable(s) also increases.

**Negative Correlation:** Two features (variables) can be negatively correlated with each other. It means that when the value of one variable increase then the value of the other variable(s) decreases.



In our dataset *Monthly Rate*, *Number of Companies Worked* and *Distance From Home* are positively correlated to Attrition.

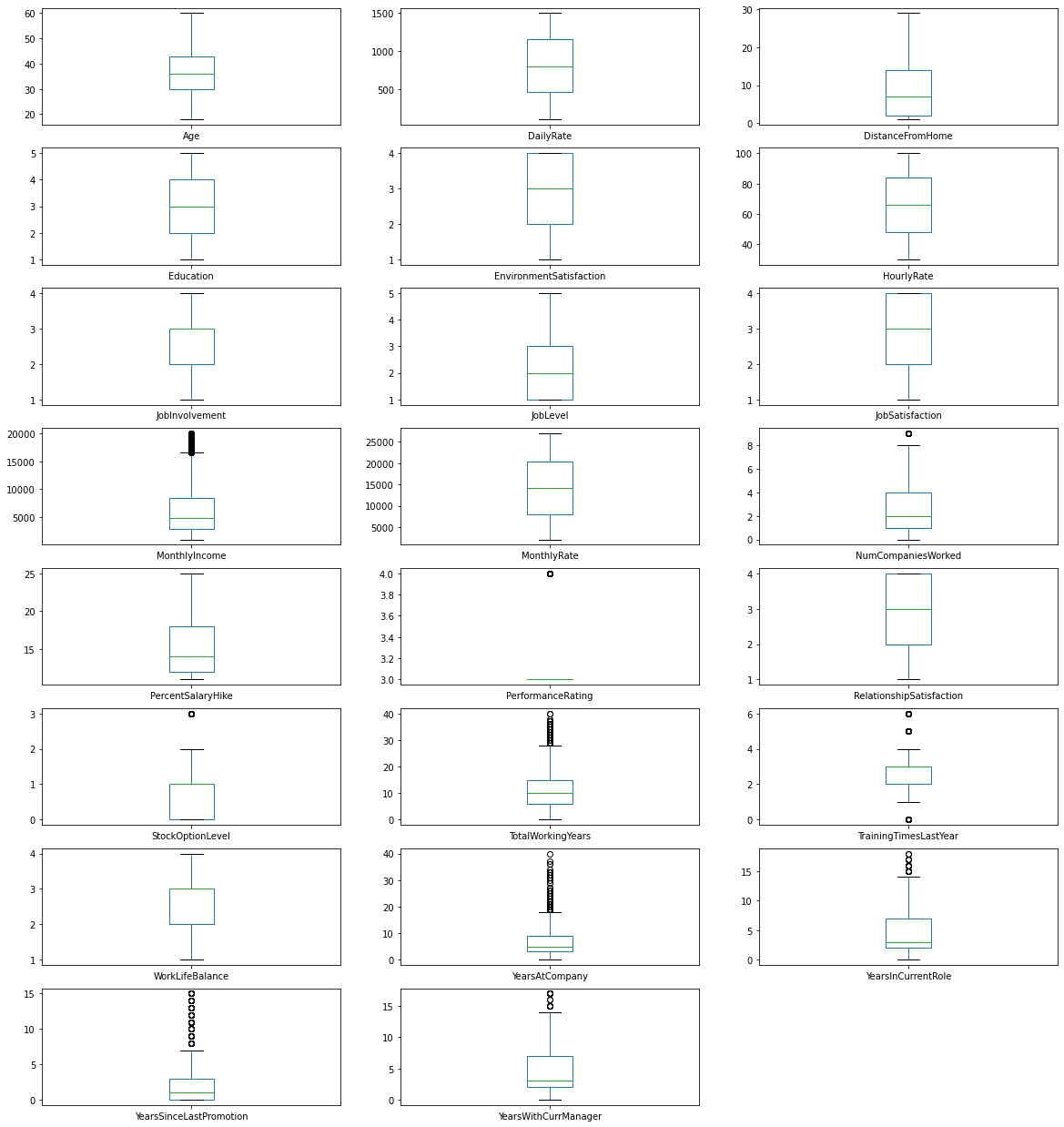
*And Total Working Years*, *Job Level*, and *Years In Current Role* are negatively correlated to Attrition.

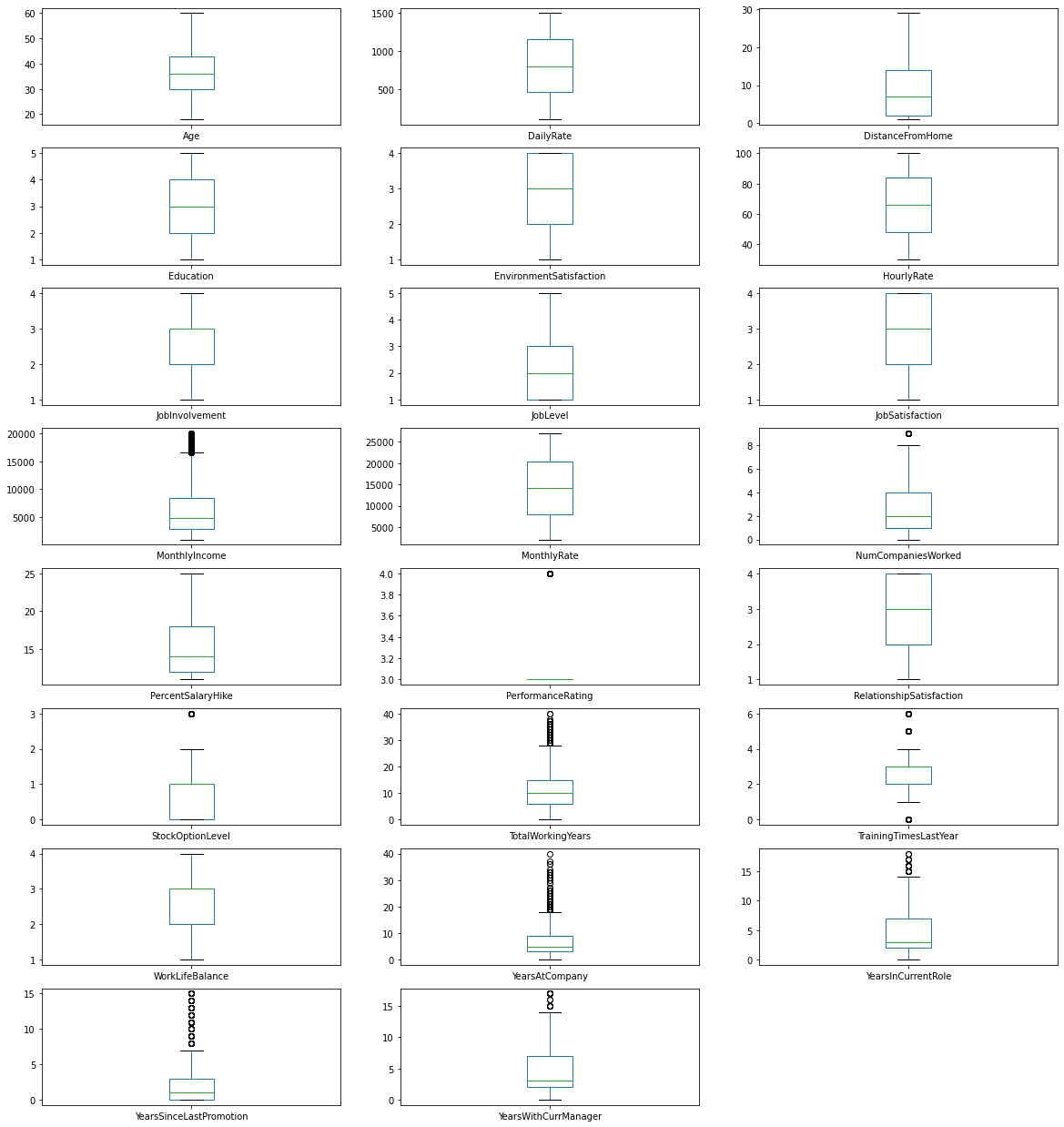
**6. Outliers:**

An **outlier** is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as outlier analysis or outlier mining.

We are using boxplot to check outliers present in our dataset.

In our dataset there are outliers in Monthly Income, Total Working Hours, Years At Company, Years Since Last Promotion, Years In Current Role, Training Time Last Year, Stock Option Level, Num Companies Worked and Performance Rating.



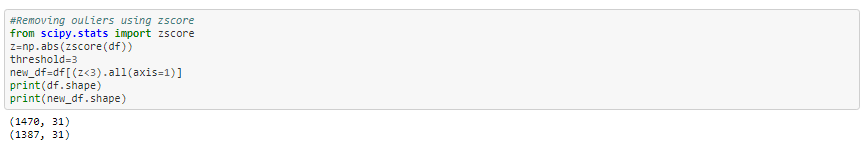


**6.1 Removing Outliers:**

We are using ZScore (also called a *standard score*) method to remove outliers.

If the z score of a data point is more than 3, it indicates that the data point is quite different from the other data points. Such a data point can be an outlier.

Below code is to remove outliers from dataset



Initially, we have 1470 rows, after removing outliers from the dataset 13 rows have been removed (1387 rows).

**7. Skewness:**

Skewness is the measure of how much the probability distribution of a random variable deviates from the normal distribution.

We are reducing skewness in this dataset using log method, whenever a column reaches 0.55 it will use log method to reduce skewness from data.



# **8. Pre-processing Pipeline**

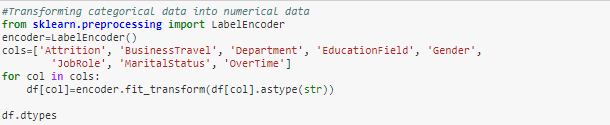
In this section, we undertake data pre-processing steps to prepare the datasets for Machine Learning algorithm implementation.

**8.1 Encoding:**

Machine Learning algorithms only have numerical values as their predictor variables, so we are using **Label Encoding** to encode categorical data into numerical data, thus all data have unique identification.

By default, Label Encoding method changes object datatype into integer datatype.

Below code is used to encode categorical data

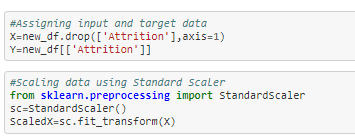


**8.2 Feature Scaling:**

We are using **StandardScaler** to perform the task of Standardization, and it follows Standard Normal Distribution. It Standardize features by removing the mean and scaling to unit variance.

Standardization of a dataset is a common requirement for many machine learning estimators.

We are standardizing input data and storing it in a new variable *ScaledX.*



## **9. Building Machine Learning Model:**

It’s time to build a Machine Learning model with our final dataset.

We are splitting data into train and test dataset because to predict the model we should have test dataset and train dataset is used to train the model.



Test size 0.33 means 33 percent of the data will be split to test dataset and remaining 67 percent data is for train dataset.

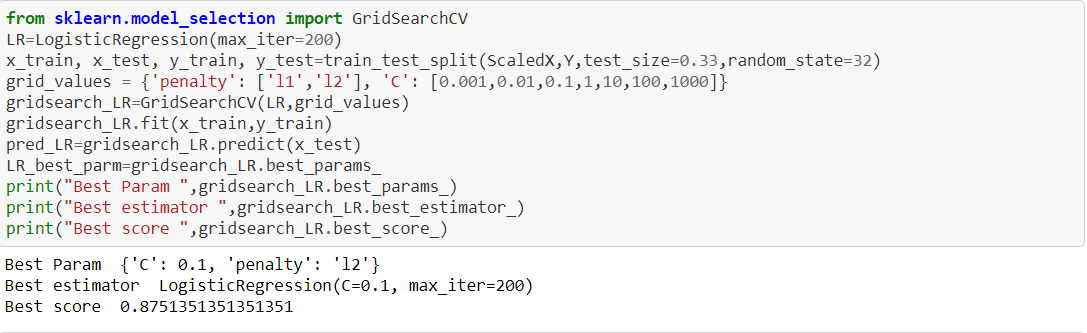
We have best random state which is 32.

**Random state**: Random state is basically used for reproducing your problem the same every time it is run. If you do not use a random state in train\_test\_split, every time you do the split you might get a different set of train and test data points.

After splitting, we are checking which model gives good accuracy score. **Logistic Regression** is the best model with accuracy score of 0.886 at random state 32.

**10. Hyperparameter Tuning:**

Now we are performing hyperparameter tuning using GridSearchCV for Logistic Regression model to get best parameter. And we got **LogisticRegression(C=0.1, max\_iter=200)** as the best estimator.



Then using best estimator, we are finding cross-validation score and accuracy score.



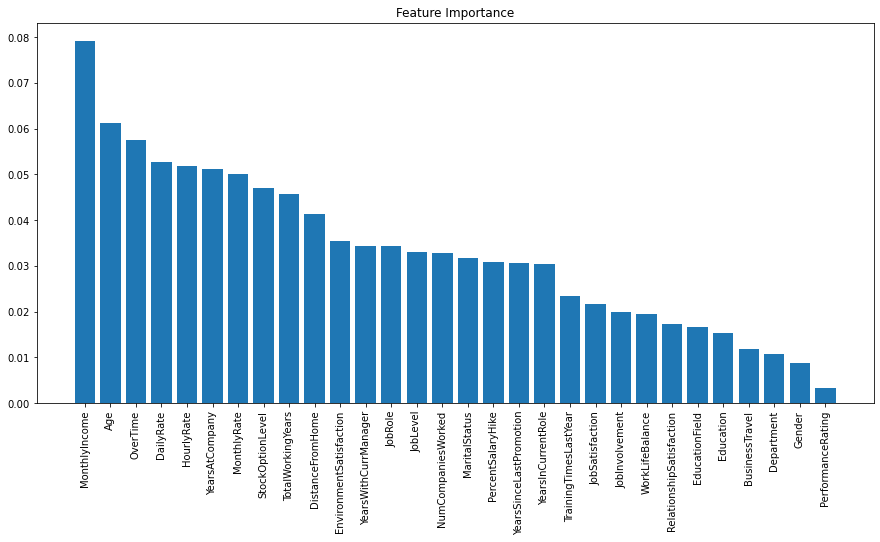
**11. Random Forest Classifier:**

**Why we are using Random Forest Classifier?**

The Random forest works on Bagging principle. It is an ensemble of Decision Trees. The bagging method is used to increases the overall results by combining weak models.

One of the best feature Random forest model has it provides the importance of variables/features in the data/model. For this HR Analytics problem, we are interested in knowing which feature/factor contribute the most in the Attrition and Random Forest is one function can give us this information.

This is the reason why we have used Random Forest.



Above visualization shows the Top most important indicators that contributes for employee attrition.

Top 10 indicators for employee attrition are:

1. Monthly income
2. Age
3. Over Time
4. Daily Rate
5. Hourly Rate
6. Years at Company
7. Monthly Rate
8. Stock Option Level
9. Total Working Years
10. Distance from Home

**12. Conclusion:**

we saw Data is important in Human Resource department (actually in most of the places, it is important). In this project, we used Machine Learning to predict whether a person is leaving a company or not.

After importing the data, we did EDA process and removed unwanted data. Analyzed data using plots. Removed outliers using zscore method and Scaled data using Standard Scaler.

Logistic Regression is the best model and scores are:

**CV score** is 0.8802956302956303  
**Accuracy score** is 0.8908296943231441.

Random Classifier shows indicators(data) of an employee leaving company. And learned how it can be very advantageous over other available machine learning algorithm.

Most of all we found factors which are most important to employees and if are not fulfilled might lead to Attrition. One of the best feature Random forest model has it provides the importance of variables/features in the data/model.

For this HR Analytics problem, Monthly Income and Age are top two indicators that contribute the most in Attrition of employees.